Visual Speech Mitigates the Influence of Speech Rate on Speech Perception

Research Thesis

Presented in partial fulfillment of the requirements for graduation *with research distinction* in Psychology in the undergraduate colleges of The Ohio State University

by

Lauren Garner

The Ohio State University

April 2020

Abstract

Sentence perception can be affected by manipulating speech rate. Dilley and Pitt (2010) found that small function words (e.g. “or”, “are”) could be made to disappear in a sentence by slowing down the rate of the surrounding context. When the target region containing the function word is shorter in comparison to the context rate, perceived target word boundaries disappear. Expanding on this finding, the current study tested if the addition of visual cues could reverse the effect and cause function word reports to increase despite the slowed context. With the inclusion of visual cues that clearly show the function word, the proportion of function word reports generally increased among slowed-context sentences compared to sentences without the visual articulation of a function word. Normal-rate sentences displayed consistently high function word reports regardless of whether the visual function word was present or absent. The results suggest that speech rate effects can be negated with the integration of conflicting visual speech cues. While rate cues aid in determining word onset and perception for casual speech, their effect can be overridden with additional visual input.

Visual Speech Mitigates the Influence of Speech Rate on Speech Perception

During conversational speech, information is communicated and received in a seemingly straightforward fashion. However, our perception of speech is actually siphoned from a variety of different signals: Background noise, a continuous string of speech sounds, and additional sensory input are all presented at the same time. Yet the human mind can filter through them and interpret meaning and linguistic regularity, allowing comprehension from an otherwise overwhelming set of signals.

This process of segmenting out meaningful information from a speech stream is possible through the use of various cues from the speech signal. Vocal productions by a talker include differences in rhythm, intonation and speed--all of which play integral parts in speech perception and comprehension. Manipulation of these cues can directly affect phonemic perception (Nooteboom, 1981). The duration of various speech sounds can additionally clue in listeners to the location of word boundaries through recognition of typical speech patterns (Shatzman and McQueen, 2006). The final product of perception is then the perceived phoneme or morpheme, intuitively derived from the auditory signal and its corresponding auxiliary cues (Öhman, 1975).

One of the more prominent cues used in speech perception is speech rate. We know that in noisy environments, speech rate differences between auditory streams aid in attention, and can helps listeners separate signals from additional noise. If someone is listening to a conversation at

one speed, and they suddenly hear an auditory stream at a different speed from somewhere else in the room, they're able to disconnect the two from each other due to the differences in speech rate. Speech rate cues also cover gaps in auditory information. In conversational speech, some sounds are dropped or lost to noise (e.g. *or*, *are*, *and*, etc.)—but rate cues can indicate intended phonemes, covering spectral holes and strengthening perception (Koreman, 2006). This is why listeners can still understand conversational speech, despite the words not being enunciated as clearly. Speech rate cues have the ability to assist in the interpretation of ambiguous auditory cues (Scharenborg, 2009), and would typically increase coherence of a speech stream. However, the manipulation of rate cues can sometimes do the opposite, inhibiting the perception of spoken morphemes. As the information derived from rate cues extends an expected syntactical pattern, a masking effect can be created by changing the larger context of an auditory signal, and diminishing the presence of a smaller portion.

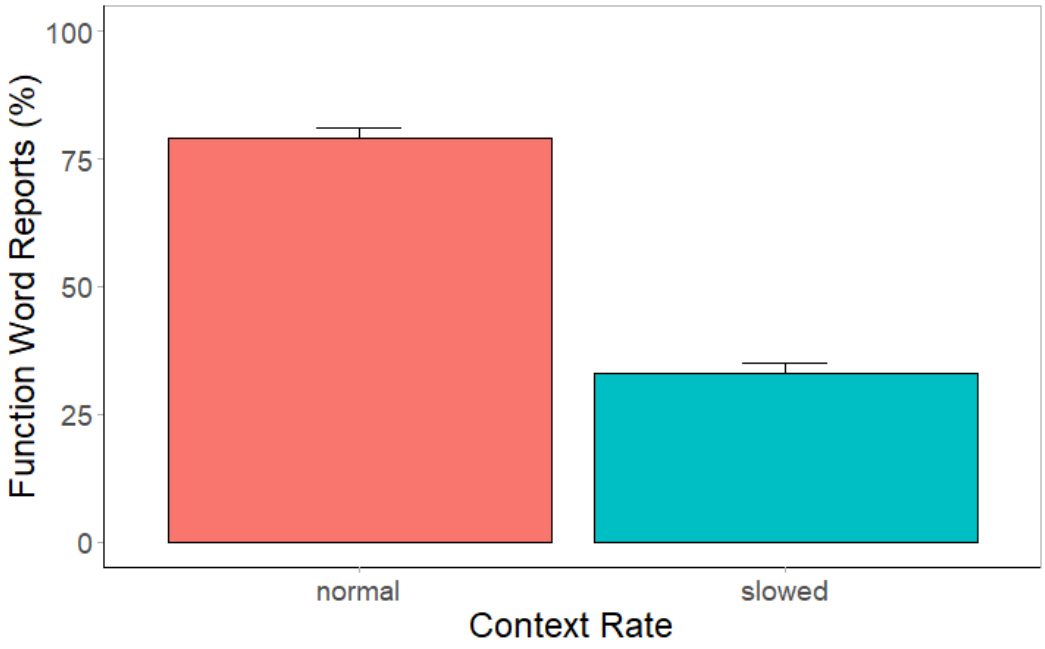
In a study conducted by Dilley and Pitt (2010), specific rate manipulations to certain English phrases were found to induce a unique perceptual phenomenon. In casual English speech, small function words (e.g. "or", "are", "and", etc.) tend to be reduced. They are not enunciated as clearly as subject words and can often be implied from context. Dilley and Pitt found that by encasing these function words in the context of a longer phrase, and lengthening the duration of that context without altering the length of that word, the function words could perceptually "disappear". This disappearing word effect (DWE) could only happen in specific scenarios, however. A function word would have to be located toward the middle of a sentence, with a context word that could be coarticulated with the function word preceding it. For example, the word "leisure" could precede the function word "or". Since the casually shortened "or" is

already fairly ambiguous, it could then appear to just be an extension of the final /r/ sound of the word “leisure” (See Figure 2).

Dilley and Pitt created a number of phrases that fit these requirements, and provided two auditory versions of the sentences. The first were normal unaltered versions of the sentences, which were recorded at a conversational pace, with present—but naturally shortened—function words. Nothing was manipulated in these sentences. The other version of the sentences involved lengthening the context of the sentences around the target function words. In these slowed-context sentences, the targets were left untouched, kept at the same conversational shortened length from the normal-rate sentences. The rest of the sentences was artificially lengthened, creating the environment necessary for rate effects to induce DWE. Participants were then instructed to listen to the sentences and type back what they heard. They found that function words were reported about 50% less for slowed-context sentences than for normal-rate sentences (Figure 1). They attributed this phenomenon to the generalized-rate normalization hypothesis, which suggests that perception of phonological and morphological units is dependent on the speech rate of the uttered sentences and surrounding context. When people communicate with each other, there are a lot of individual differences between speech quality. One of those differences is the speed at which that individual talks. Because individual speech rates vary so heavily between people, it’s hypothesized that listeners have the ability to adjust to a general speech rate for a talker, in order to efficiently determine morpheme boundaries amidst vastly different signals. This effectively sets a pace for the auditory signal, and many sudden changes can be assimilated in.

Figure 1

Dilley and Pitt DWE Results



Note. Dilley and Pitt (2010) results (N=41), displayed in proportion of function word reports between normal-rate sentences and slowed-context sentences. Normal-rate sentences showed 79% function word reports, while slowed-context showed 33% reports.

Dilley and Pitt (2010) provided valuable insight toward how word boundaries and word onsets are determined, and how speech rate cues can aid in the understanding of ambiguous auditory signals. Further studies have been conducted on additional aspects of the top-down

processing used in this phenomenon. Examinations of rhythm cues distant from the target in sentences and their effects on the phenomenon (Morrill et al., 2014) proved fruitful, finding rhythms that were predictive of the target function word region elicited more function word reports than non-associated rhythms. These rhythms were pitch patterns that varied dependent on the presence of a function word in a sentence—variations of repeating ternary or binary pitch patterns. One study explored the extent of the timescales for rate effects used to determine word perception, finding that longer participation in an auditory event, or longer time spent listening and acclimating to a conversation, increased the effects of speech rate over time (Baese-Berk et. al., 2014). Another study examined how speech rate effects functioned in noisy environments, and how dependence on them changed as the auditory signal was manipulated (Gibson et. al., 2013). They found that the rate effects in noisy environments created the perception of more plausible sentences, filling in gaps and reinterpreting signals based on the most likely outcome for that sentence.

All of these studies probed the top-down rate effect nuances on the disappearing word effect, and found that the DWE was robust in multiple manipulations. However, one source of additional speech information has not yet been explored in the context of the DWE. In many realistic scenarios, a listener is provided with more than just auditory stimuli—they also have access to visual information coming from the speaker in a conversation. In addition to manipulating the top-down nature of the rate effects in the auditory signal, we could also examine the role of visual information in interpreting what a talker said. Introducing a visual stimulus along with the auditory speech rate manipulation from Dilley and Pitt's effect could influence perception of the function word.

There are numerous studies regarding the influence of visual cues on speech perception, with perhaps one of the more notable being the McGurk Illusion (McGurk, 1976). This study manipulated participants' perception of a spoken morpheme (e.g. "ba") by pairing it with an incongruous visual stimulus of a different morpheme being produced (e.g. "ga"). The "ba" sound participants heard was influenced by the visual speaker mouthing "ga", which in turn caused the participants to approximate a different sound as the perceived stimulus. They reported hearing "da" instead, after receiving visual input from a velar production and auditory input from a bilabial production. The final reported sound fell in the middle, as an alveolar production.

We know AV integration isn't completely impervious, as seen in the study by Windmann (2004), which showed that AV integrated perceptions, previously thought to be robust against other cues, were also subject to the tendencies expressed by ambiguous phonemes. The interpretation of ambiguous stimuli is often dependent on lexical-semantic context and other additional cues (like speech rate), which influences their perception. This influence in perception, due to cognitive functions, was found to also manipulate perceptions made from AV cues in this experiment. This was studied by using the McGurk effect in fuller morphological and semantic contexts--in real word studies and sentence content expectations. The success of the illusion was found to be vulnerable to semantic expectations—if the word being targeted was in a sentence location that made logical sense vs a semantically improbable position, the McGurk Effect would differ in intensity. The illusion would break down when the word was located somewhere illogical, as opposed to a typical, semantically acceptable location. For example, if the word using the McGurk Effect was "teeth", the effect would be stronger if it was in a sentence where "teeth" was expected, as opposed to "*On an orbit in space you find Mars, the*

Earth and every other TEETH". This argues that AV strength can be diluted by additional cognitive factors. But it still maintains a robust relationship in the face of other cues.

The addition of visual information to an auditory stream can cover major gaps in perception and increase coherence in very noisy situations. Even in distortion, visual input can aid in the interpretation of ambiguous stimuli (Eg et. al., 2015). It aids in conversational attention, and allows a listener to focus on an auditory source, even if the environment is full of additional interruptions (Shahin and Miller, 2009). As visual input directs the listener to the source, and keeps the attention fixed there, it can become highly suggestive if the listener believes the visual and the auditory sources to be one and the same. This can be seen in the ventriloquism effect (Jack and Thurlow, 1973). When the true audio source is hidden, and the observer is provided with a visual signal that matches the temporal cues of the audio, they can attribute the audio source to be that of the unassociated visual signal. Therefore, when the listener believes both signals originate from the same source, their perception of unbroken clear speech originating from the visual source is encouraged and is stronger than if there were highly disparate signals.

The previous studies demonstrate that audiovisual integration is a robust and automatic process, and that visual cues can significantly impact interpretation of an auditory speech signal. The addition of visual cues in the DWE could impact perception of function words. If listeners are provided with contradictory visual cues indicating the presence of a function word, along with the slowed-context sentences, would they report more function words than slowed-context sentences with visually absent function word cues? Because the slowed-context sentences make it seem like the target word has disappeared, despite it still being present, indicating visually that it is still there may induce its perception once again. Not indicating the function word visually

should further strengthen the DWE effect in turn, since it would discourage function word perception during AV integration. The DWE occurs because participants disambiguate the short function word signal with rate effect cues—but if additional visual stimulus was present as another form of information to override ambiguity, would participants prioritize that form of influence over speech rate effects?

I wanted to examine two sets of conditions and their effects on each other—through the slowed-context and normal-rate sentences, and through visually absent function words and visually present function words. I hypothesized that pairing slowed-context sentences with clearly visible function words should increase function word reports in comparison to slowed-context sentences with visually absent function words. The DWE should be manipulated depending on presence or absence of a visual function word, even if speech rate effects would typically cause the word to disappear. The current study included two experiments to examine this hypothesis. The first experiment manipulated the visual input for only slowed-context sentences. The second experiment acted as a control, manipulating the visual input for the less ambiguous normal-rate speech. This would prove that the AV influence would only be probable in ambiguous scenarios like DWE, and any differences would be due to our manipulation of visual target presence or absence.

Experiment 1

The first experiment manipulated visual target word presence for solely slowed-context sentences. All target sentences exhibited DWE in this slowed-context form, and were given two conditions: visually present and visually absent function word videos. I predicted that function word reports would increase with a visually present target versus a visually absent target,

indicating a prioritization of AV cues over rate effect cues when perceiving the ambiguous auditory function words.

Method

Participants. Participants were self-identified native English-speaking undergraduate students from The Ohio State University (N=27), all with normal or corrected vision and self-reported normal hearing. They were compensated with course credit through Ohio State's REP program for completing the experiment.

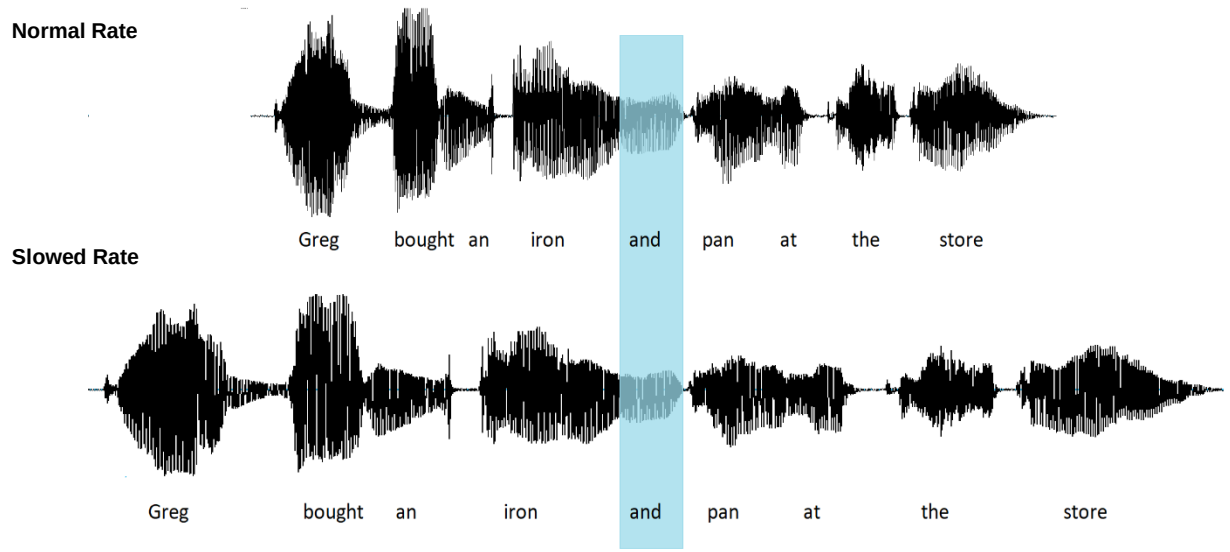
Stimuli. Sixteen sentences were constructed for three function word categories: *and, or*, and *are*. These sentences ranged from about 8-12 words in length, and contained a context word before the target function word that encouraged phonetic coarticulation (e.g. *leisure or, iron and, butter are*).The sentences made grammatical sense with or without the function word present in the sentence. This can be exemplified in the sentence *Greg bought an iron (and) pan at the store*. Whether or not the function word is perceived, the sentence still makes sense to a typical English speaker. Seventeen filler sentences were created as well, with similar length to the target sentences (See Appendix A). These were mixed in with the target sentences to pad the trials and avoid participant bias. The experiment was then able to be framed as a memory test.

The stimuli were recorded auditorily by a native English speaker with an Ohio Midwestern dialect. A TASCAM HD-P2 audio recorder and an EV N/D 308 Dynamic Instrument microphone were used to record the auditory stimuli, recording at a sampling rate of 48,000 Hz. Video stimuli were recorded using a SONY 4K Handycam video camera, recording at 59.95fps.

The audio files of the target sentences were then manipulated in Praat, changing the rate at which the context of each sentence was played (Boersma, 2001). The target function word region was first selected by establishing the leftmost boundary at the final syllable of the preceding context word, and the rightmost boundary at the first phoneme of the following context word. For the sentence *Greg bought an iron and pan at the store*, the target region would be [-*ron and p*-]. Leaving this target region untouched, the surrounding context of the sentence was slowed down 1.5 times, using the Lengthen tool on Praat. This means the function word target would be identical to the normal-rate sentence, no matter the manipulation (Figure 2). These slowed-context stimuli were tested and replicated the effect in the study by Dilley and Pitt (2010), causing lower function word reports relative to the normal-rate sentences (see Appendix B). The filler sentences were slowed in entirety by the same 1.5 factor, to remain consistent with the target sentences.

Figure 2

Slowed-Context Manipulation



Note. An image depicting the waveforms of the target sentence *Greg bought an iron and pan at the store*. The top depicts the normal spoken rate, the bottom depicts the slowed-context rate. In both conditions, the target region remained the normal spoken rate, not manipulated in any of the experiments.

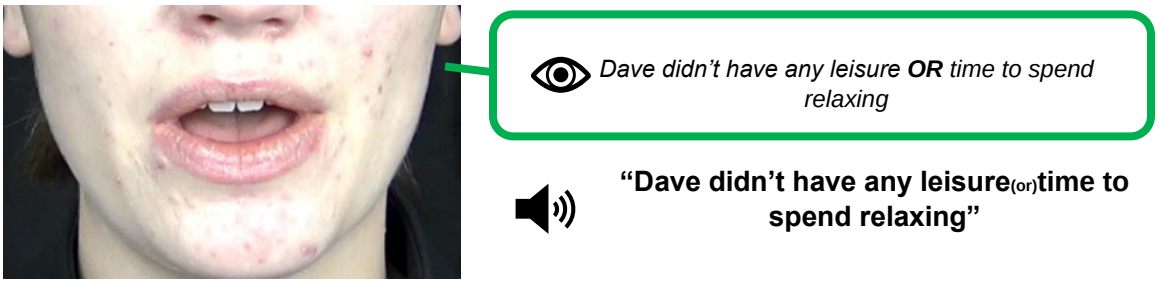
The slowed-context sentences were then paired with a video of a person mouthing along with the sentence. An individual was recorded speaking along with the slowed-context audio clips, enunciating the words clearly. There were two conditions for the video stimuli for each slowed sentence. The first condition was the sentence being produced without the function word visually present. If the slowed-context sentence was *Greg bought and iron (and) pan at the store*, then the video would clearly mouth *Greg bought an iron pan at the store.* The other condition inserted the function word into the sentence visually. For the same slowed-context sentence, the visual would clearly mouth *Greg bought an iron and pan at the store* (Figure 3). Video stimuli were also recorded for the slowed filler sentences and did not contain a visual manipulation. The videos were recorded separately from the audio and were edited in Adobe Premiere. The videos were cropped to show the tip of the nose to the bottom of the chin. This was to avoid potential distractors from facial expressions and other potential cues.

Figure 3

AV Stimulus

A)

B)



Note. Slowed-context stimuli were paired with a visually present function word video, and a visually absent function word video. Figure 3a depicts a visually absent function word, and Figure 3b depicts a visually present function word.

Procedure. The experiment consisted of 31 trials, including 16 target sentences and 15 filler sentences for each participant. There was a present function word condition (present condition) and an absent function word condition (absent condition) for each of the 16 slowed-context target sentences, creating 32 unique audiovisual slowed-context target stimuli. These audiovisual target conditions were separated randomly onto two different experiment lists so that each list contained an equal number of sentences with the visual function word present or absent. Both lists contained all the slowed filler sentences. Each list started with two practice trials using filler sentences to familiarize participants with the task.

Participants were instructed to pay close attention to the talker on the screen and remember exactly what the talker said. A video stimulus was presented on each trial; after which

participants were given two alternative forced-choice options. For the target stimuli, these showed fragments of the sentence containing the preceding and following word around the target function word, with or without the function word present (e.g. *iron and pan* vs. *iron pan*). The filler sentence just had three words chosen from the sentence, and a variation of that sentence which was false. Participants were then instructed to choose the option which had been part of the sentence the speaker had said. All stimuli were presented over headphones through a custom-made, browser-based experimentation framework. A questionnaire was provided at the end of the experiment to gauge the difficulty of the task, as well as language experience.

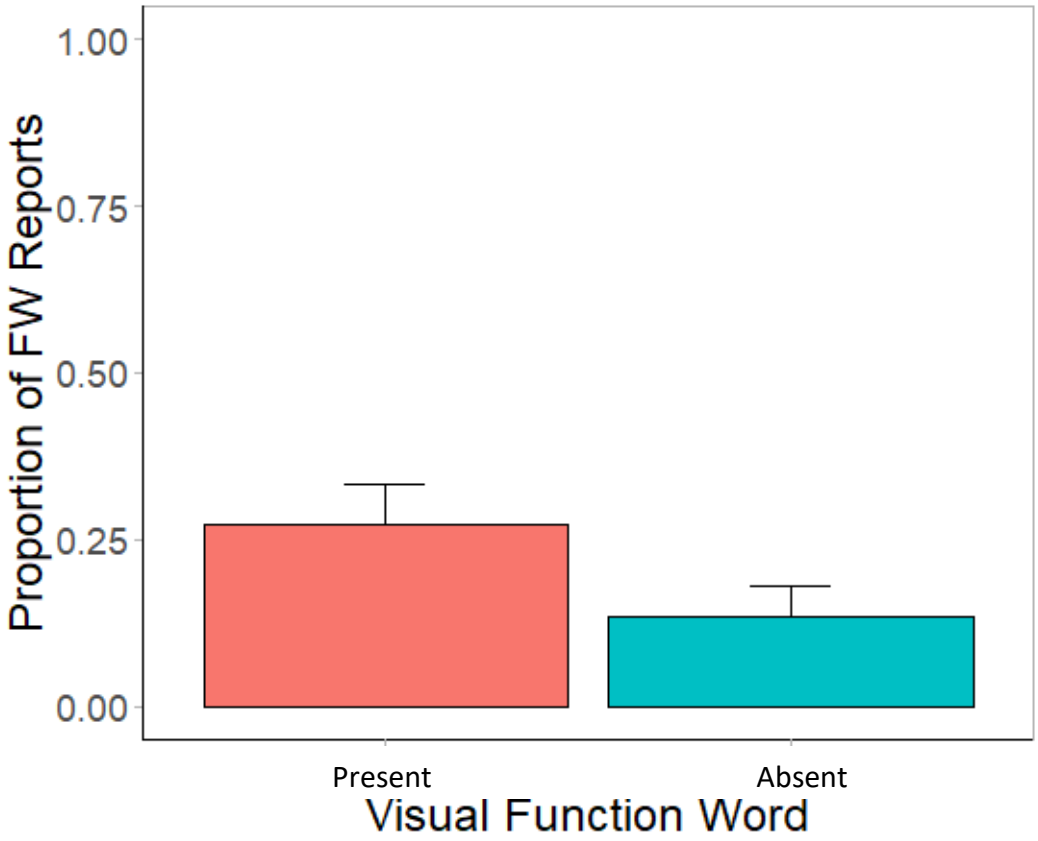
Results and Discussion

The forced choice selections for target sentences were scored. Those responses where a function word was perceived were scored as 1, while the absent responses were scored as 0. These were then proportioned over the total responses for each condition, and graphed by their percentage of function word reports.

The results of Experiment 1, seen in Figure 4, show a difference between present and absent function word videos. A higher proportion of function word reports, about 27%, were associated with a present function word video, while only 13% were exhibited for trials with an absent function word video.

Figure 4

Slowed-Context Results



Note. Proportion of function words reported between visually present and visually absent function word videos for slowed-context sentences (N = 27). Error bars indicate 95% confidence intervals.

A paired samples t-test indicated a significant difference between with visually present and visually absent conditions ($t(26) = 3.483$, $p = 0.001$). These results do show a significant increase in function word reports when the function word is visually articulated versus when it is visually absent.

Both proportions were much lower than reports from the Dilley and Pitt auditory DWE experiment, however, showing no complete reversal of the effect with the addition of visual information. This may be due to the additional cognitive load from the combination of audio and

visual cues. As expressed in a study by Strand and Brown (2019), visual face cues during speech can contribute fine phonetic detail to aid with word recognition and perception, but those stimuli are actually more taxing on memory than more simplified visual cues, or solely auditory stimulus. Certain aspects of speech perception are enhanced in terms of recognition, but processing power goes up with the addition of each new stimulus type. This may have affected memory for function words towards the middle and end of the target sentences, seeing a reduction in function word reports overall in the audiovisual experiments.

The results remain consistent with the hypothesis that slowed-context sentences, which show DWE in auditory trials, could be paired with present condition videos, showing higher function word reports than those with absent condition videos. This would indicate that the additional visual cues can further manipulate morphome perception, creating a distinct difference in function word reports when presented with either a visually salient signal, or a visually absent signal. This is most likely due to the ambiguous nature of the auditory stimuli—the majority of the sentence context can be understood, being just spoken at a slower rate than normal speaking speed, but the target function word is just barely perceptible, its presence normally being inferred from the surrounding context rate clues. As soon as a contradictory visual stimulus was introduced, however, a more concrete speech cue was utilized that lessened the influence of the rate effects. The AV signal's perception took precedence over the rate effects influence on perception, and became the preferred method of disambiguating the sentence. Since both the auditory and visual cues appeared to come from the same source, the combined signal appeared less ambiguous, with visual information suggesting the function word was still present in the sentence. The ambiguity of the slowed-context sentences most likely determined whether this AV influence would occur—but I needed to clarify whether this influence took precedence every

time rate cues were used, or only in ambiguous situations. My second experiment was constructed to examine this possibility.

Experiment 2

The second experiment examined how the visual function word manipulation affects the processing of normal-rate sentences. The conditions in this experiment provided baseline performance when normal-rate stimuli are presented to listeners. I predicted that perception of normal-rate sentences would remain fairly unaffected by any manipulated visual function word presence. The function words present in the auditory stimulus would be much less ambiguous than those in the slowed-context sentences. This would potentially lower reliance on additional visual cues to interpret sentence content. Function word reports should remain high, and at similar proportions for both present and absent function word videos.

Method

Participants. Participants were self-identified English-speaking undergraduate students from The Ohio State University (N=30), all with normal or corrected vision and self-reported normal hearing. They were compensated with course credit through Ohio State's REP program for completing the experiment.

Stimuli. The 16 original target sentences were used again for the second experiment, but only sentences produced at their naturally spoken rate were used. They were paired with both present and absent function word videos, recorded in the same manner as the first experiment, paired with only the normal-rate sentences. Filler sentences were also presented at their naturally produced rate.

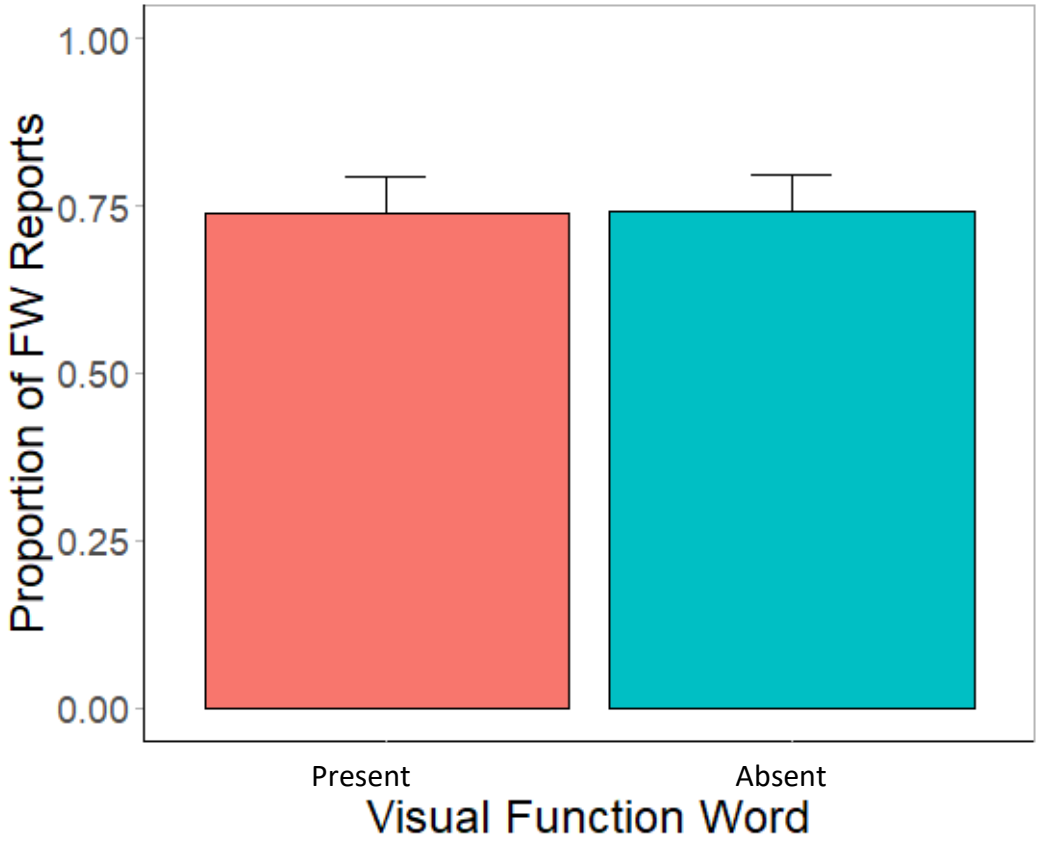
Procedure. The procedure for Experiment 2 was identical to that of Experiment 1, with the exception of using normal-rate stimuli in place of the slowed-rate stimuli.

Results and Discussion

I scored function word reports in the same way as Experiment 1, recording the forced choice responses that included function words and creating proportions for present and absent conditions. Function word reports were high for both video types among normal-rate sentences, regardless of whether the visual function word was present (.74) or absent (.74) (Figure 5). There was no significant difference found between the conditions ($t(29) = -0.111$, $p = 0.912$).

Figure 5

Normal-Rate Results



Note. Proportion of function word reported between visually present and visually absent function word videos for normal-rate sentences (N = 30). Error bars indicate 95% confidence intervals.

These results were regarded as control data, confirming AV influence and preference only in instances where the auditory signal is ambiguous. Any differences between function word reports in Experiment 2 would be due to visual manipulation during normal-rate speech, and no difference was recorded. During normal conversational speech, there was no significant AV influence. There is no added ambiguity past what is typically present in colloquial speech. The DWE was exhibited in slowed-context speech because participants were using speech rate to interpret the ambiguous region containing a function word—but when the function words are surrounded by speech spoken at the same rate, no visual cues are required to interpret what was being said. Participants can clearly hear the function words almost every time, with some variability due to the natural ambiguity of conversational speech. This would seem to indicate that audiovisual cues are prioritized and depended on more when the auditory signal isn't as clear. More information from outside the ambiguous auditory stream must be utilized in order to best interpret what is being said.

General Discussion

The purpose of this study was to examine the relationship between AV integration cues and rate effects' influence on ambiguous stimuli, to see which are prioritized in the DWE phenomenon. The results from Experiments 1 and 2 were very promising—Experiment 1 found that perception of function words in slowed DWE sentences increased when combined with present condition videos as opposed to absent condition videos. This indicated that in the

auditorily ambiguous DWE scenarios, when implicating AV integration, visual cues can lessen the effects of speech rate on perception. Experiment 2 explored the same visual manipulation as Experiment 1, but with only normal-rate sentences. These sentences were not rate manipulated, and for that reason, remain mostly unambiguous. Function word reports were high for both visually present and visually absent function word videos, since the function word always seemed to be present auditorily, and was not swayed by the presence of visual stimuli that suggested otherwise. The results from these experiments did support the original hypotheses, exhibiting significant differences in function word reports between the video types in slowed sentences, and no significant difference among normal-rate sentences.

These experiments identify the settings where audiovisual cues show dominance, providing additional perceptual information to a listener. When the auditory stream is clear and understandable, additional cues outside of rate aren't necessary to infer meaning—but when the auditory stream is ambiguous, even with the generalized context rate as a guideline, audiovisual cues hold more weight and help interpret those vague signals. Though at odds with each other in these specific experiments, the interaction between audiovisual and speech rate actually suggests that both types of cues can aid speech perception, but take precedent in different settings. During normal conversational speech, speech rate cues are used to determine probable word boundaries and aid in adapting to a talker's unique speed. Additionally, in auditorily ambiguous situations, rate effects can be used to try to determine the most plausible meaning from vague signals. But, in those ambiguous situations, there is a preference for audiovisual influence on perception over speech rate influence, perhaps because visual information typically remains unaltered during auditory holes—It can supplement information if missed. Audiovisual integration did not influence function word perception during normal-rate speech. This indicates no additional

necessary dependence on audiovisual cues when the speech signal was clear. If both speech rate and visual cues are aligned, then understanding of an auditory signal is enhanced. If the cues are contradictory to each other, then the setting could determine whether the perceptual interpretation remains influential.

The current study sought to examine the DWE from Dilley and Pitt (2010) under more realistic conditions, with the addition of visual input when perceiving slowed-rate sentences. This was because the original DWE experiment took place in a solely auditory setting, while a good number of real-life situations involve being able to see the talker as well as hear them. It is important to determine whether any additional signals that are common to real world scenarios could influence the phenomenon—and my study did yield such results, finding that visual cues in tandem with the auditory signal can lower the robustness of speech rate effect on perception. It combined the efforts of previous AV studies—examining the influential relationship between visual and auditory stimuli—with information about the similar relationship between speech rate and speech perception. Both visual cues and speech rate are strong influences on speech perception, but they were found to operate in different positions of prominence depending on the scenario. When the speech signal is ambiguous, as in the slowed-context sentences, visual cues have a stronger effect on perception. When it is less ambiguous, speech rate is a more robust cue for word boundaries. They were found not to be opposites of each other, but different types of tools or mechanisms that maximize the efficiency of speech perception.

One of the more interesting questions that arose from these experiments was the fact that slowed-context, visually-present function word trials were unable to fully reverse the DWE for participants. Function word reports were overall much lower with slowed-context sentences in Experiment 1 compared to the slowed-context condition in the original Dilley & Pitt (2010) study.

Understanding the amount of conflicting information presented to participants may be a key proponent in eliciting greater reports in future studies. Finding a way to reduce the cognitive load on memory during AV trials is an imperative next step. Since participants are processing both slowed audio and visual face stimuli, the efficiency during perception is much lower than with simpler models of visual input, or normal-rate auditory stimuli. Participants are being exposed to multiple streams of information, including auditory and visual temporal differences, speech quality, visual face details, etc. Layering audio and visual stimulus creates more details to process during a memory task. Additionally, we found over the course of the experiments that the individual sentences used varied in terms of strength of the DWE and influence of visual cues. All of the sentences used during the experiments were tested using the Dilley and Pitt (2010) DWE method (see Appendix B), to make sure the function word proportions were the same between our stimuli and theirs. In the AV pilot trials, however, there were differences between "or", "are" and "and" that could be attributed to visual salience issues. The function word "or" was very visually salient, opening the mouth in a larger, round shape. The other two function words could tend to get lost in the rest of the phonemic shapes in the context. It would be beneficial to examine the differences between function word visual salience in future studies. These were adjusted to increase visibility for the final experiment trials.

The experiments did answer a number of different questions about external cues and how they play into speech segmentation—depicting how syntactical expectation and grammatical patterns can be influenced by more innate, physical cues through audiovisual integration. These supposed bottom-up audiovisual cues in tandem with top-down temporal cues all aid in the segmentation of the ongoing speech stream. However, there may be another element aiding in the effect depicted in Experiment 1—audiovisual cues, though typically attributed to be a more

bottom-up cue, can actually display some more top-down linguistic aspects. As seen in a study by Wang et. al. (2008), audiovisual perception can actually differ between individuals dependent on their linguistic knowledge and language experience when identifying speech sounds. Similarly to rate effects, the effectiveness of the mechanism is based on the understanding a listener has about the tendencies and patterns of a language—what they expect to see based on their knowledge and familiarity. Future research may attempt to isolate whether audiovisual cues are acting as a bottom-up or a top-down influence on speech perception in Experiment 1.

Examining the relationship between audiovisual cues and speech rate in ambiguous auditory signals has displayed a mitigating effect present on rate-based perception in these settings—their influence diminishes as audiovisual integration effects are prioritized, demonstrating the strength that visual cues have in interpreting ambiguous speech. In naturally unambiguous environments, such as conversational speech, the precedent moves the opposite direction, favoring speech rate as a dominant cue to determine the correct segmentation of the auditory stream. Both can be used in those vague environments, but at different levels of effectiveness. It is still maintained that the strongest perceptual interpretation comes from maximizing use of both types of cues during ambiguous speech.

Appendix A

Stimulus List

Targets

1. Annie would only pay in silver (or) coins to avoid fraud
2. Dave didn’t have any leisure (or) time to spend relaxing
3. The music drifted up from the floor (or) vents very softly

4. She had timeout in the corner (or) closet without any dinner
5. Sam had shut off the breaker (or) switch immediately
6. The surgery required a new donor (or) heart for success

1. Greg bought an iron (and) pan at the store
2. Frank cooked the salmon (and) burgers on the grill
3. Jared works at the tavern (and) bar every Saturday
4. Kelly served bacon (and) jam with breakfast
5. He makes deliveries at seven (and) ten every day

1. Chris sees the milk and sugar (are) on the table
2. The principal and teacher (are) like cats and dogs
3. The cookbook showed flour and butter (are) in the recipe
4. Ian knows the cook and baker (are) in the kitchen
5. Reports indicated lightning and thunder (are) in the forecast

Fillers

1. Annie wanted to see a very funny movie with her friends
2. It cost a lot to fix a broken saxophone at the shop
3. Peter built a large cardboard fort in his backyard
4. These are our old coins from Ancient Greece and Rome
5. The soldier fought in many battles during the war
6. Lucy didn't want him to buy an expensive boat
7. The mother didn't want her to son to watch movies
8. Scientists have tried to make a city on the moon
9. Steve baked a peanut-butter pie without using vanilla or nuts
10. The captain was only after her gold coins and jewels
11. The girl believed she could conquer her fear with courage
12. Carla bought flour to batter her fish for the picnic
13. Sarah had to look under her cabinets for the spotted cat
14. The dancer was prepared before her practice that evening
15. Many students want to travel the world while in school
16. It's rare to see a white flamingo outside of the zoo
17. The pancakes were covered in a thick syrup and butter

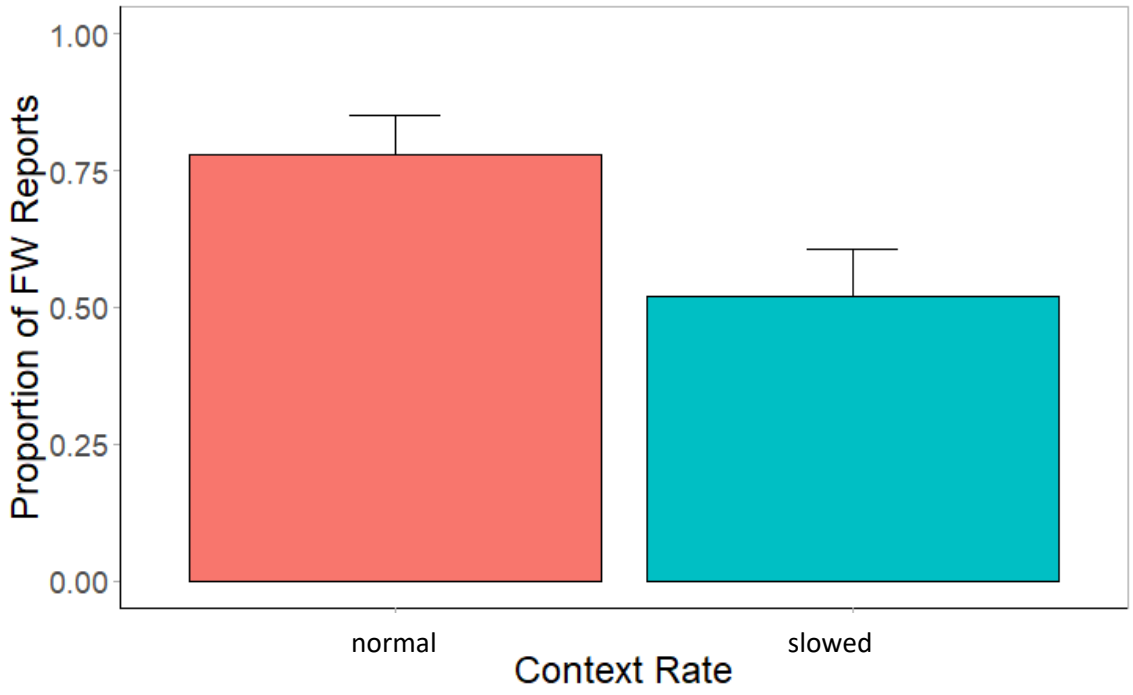
Appendix B

Auditory Pilot Results

All target sentences were piloted with the DWE manipulation from Dilley and Pitt (2010) to guarantee they elicited the same effect as the original Dilley and Pitt stimuli.

Figure 6

Auditory Trial Results



Note. The graph displays my pilot results (N=21). We found about a 75% function word report for normal-rate sentences, and about a 50% report for slowed sentences, compared to the original 79% and 33% for Dilley and Pitt.

The experiment prompted participants to listen to a sentence stimulus, then type the sentence they heard in a free response format. Function word reports were manually scored.

Acknowledgements

I would like to thank Julia Bencko, Zhiyu Bu, Yazhao Huang, Jena Jackson, Leah Marek, and Zhixin (Annie) Zhou for their help with data collection over the course of the experiments.

References

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-Term Temporal Tracking of Speech Rate Affects Spoken-Word Recognition. *Psychological Science*, *25*(8), 1546–1553. https://doi.org/10.1177/0956797614533705

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1–11. https://doi.org/10.1080/09541440601125623

BOERSMA, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, *5*(9), 341–345.

Brancazio, L. (20040524). *Lexical Influences in Audiovisual Speech Perception.* Journal of Experimental Psychology: Human Perception and Performance. https://doi.org/10.1037/0096-1523.30.3.445

Dilley, L. C., & Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. *Psychological Science*, *21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743

Eg, R., Griwodz, C., Halvorsen, P., & Behne, D. (2015). Audiovisual robustness: Exploring

perceptual tolerance to asynchrony and quality distortion. *Multimedia Tools and Applications*, *74*(2), 345–365. https://doi.org/10.1007/s11042-014-2136-6

Erber, N. P. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. https://doi.org/10.1044/jshr.1202.423

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Jack, C. E., & Thurlow, W. R. (1973). Effects of Degree of Visual Association and Angle of Displacement on the "Ventriloquism" Effect. *Perceptual and Motor Skills*, *37*(3), 967–979. https://doi.org/10.1177/003151257303700360

Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *The Journal of the Acoustical Society of America*, *119*(1), 582–596. https://doi.org/10.1121/1.2133436

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. *Cognition*, *131*(1), 69–74. https://doi.org/10.1016/j.cognition.2013.12.006

Nooteboom, S. G. (1981). Speech Rate and Segmental Perception or the Role of Words in Phoneme Identification. In T. Myers, J. Laver, & J. Anderson (Eds.), *Advances in Psychology* (Vol. 7, pp. 143–150). North-Holland. https://doi.org/10.1016/S0166-

4115(08)60188-0

Öhman, S. (1975). What is it that we perceive when we perceive speech? In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception* (Vol. 11, pp. 36–47). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-81000-8_3

O'Neill, J. J. (1954). Contributions Of The Visual Components Of Oral Symbols To Speech Comprehension. *Journal of Speech and Hearing Disorders*, *19*(4), 429–439. https://doi.org/10.1044/jshd.1904.429

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 621–637. https://doi.org/10.1037/0096-1523.4.4.621

Scharenborg, O. (n.d.). *Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands*. 5.

Shahin, A. J., & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *The Journal of the Acoustical Society of America*, *125*(3), 1744–1750. https://doi.org/10.1121/1.3075576

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, *68*(1), 1–16. https://doi.org/10.3758/BF03193651

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport*, *12*(1), 7–10.

Strand, J. F., Brown, V. A., & Barbour, D. L. (2019). Talking points: A modulating circle reduces listening effort without improving speech recognition. *Psychonomic Bulletin &*

Review, *26*(1), 291–297. https://doi.org/10.3758/s13423-018-1489-7

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–1726. https://doi.org/10.1121/1.2956483

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*(2), 212–230. https://doi.org/10.1016/j.jml.2003.10.001